Multi-task Learning for Macromolecule Classification, Segmentation and Structural Recovery in Cryo-Tomography – Supplemental Document

S1 Dataset

S1.1 Simulated subtomograms from known structures

For a persuasive assessment of the approach, we generated realistically simulated tomograms with known structures of macromolecular complexes by simulating the actual tomographic image reconstruction process. It is similar to previous works [1].

Specifically, 22 distinct macromolecular complexes (Tab. Supplementary S1) are chosen from the Protein Databank (PDB) [2] for experiments. Each simulated tomograms of $600 \times 600 \times 300$ voxels contains 10000 randomly distributed macromolecular complexes with a tilt angle range $\pm 60^\circ$. Given the true position of these macromolecules inside tomograms, we extracted the subvolumes of 40^3 voxels centering on these positions as input to our model. Removing those subtomograms outside the boundary of tomograms, we finally collected 3205 simulated subtomograms of 22 structural classes for each dataset. Datasets A, B have SNR of 0.06, and 0.01 respectively. Fig. 2a shows examples of 2D slices of subtomograms extracted from a simulated tomogram.

S1.2 Experimental tomograms

A ribosome dataset of 400 subtomograms were extracted from a tomogram of primary rat neuron culture [3]. The tomogram was captured from tilt angle -50° to $+70^\circ$. It was then binned twice to a voxel size of 1.368 nm. Subtomograms of size 40^3 were extracted from the tomogram using Difference of Gaussian particle picking method [1] and coarsely filtered by a convolutional autoencoder [4]. Template search was applied to select the top 1000 subtomograms with highest structural correlation with the ribosome template. We manually inspected the 1000 subtomograms, and filtered out 141 of them which contained obvious non-ribosome structure such as fiducial. To prevent class imbalance problem, we randomly select 400 ribosome subtomograms from the 859 filtered subtomograms.

Furthermore, DSM-Net was tested on a dataset consisting of 386 single capped proteasome subtomograms extracted from a tomogram of rat neuron with expression of poly-GA aggregate [3]. All subtomorgams were two times binned to size 40^3 (voxel size: 1.368 nm). The tilt angle range was -50° to $+70^\circ$.

Overall, 400 ribosome and 386 single capped proteasome subtomograms are combined and shuffled, named Dataset C. The segmentation and density map ground truth were prepared by aligning the corresponding structural template (PDB ID: 5T2C and 5MPA).

PDB ID	Macromolecular Complex	
1A1S	Ornithine carbamoyltransferase	
1BXR	Carbamoyl phosphate synthetase	
1EQR	Aspartyl-TRNA synthetase	
1F1B	E. coli asparate transcarbamoylase P268A	
1FNT	Yeast 20S proteasome with activator PA26	
1GYT	E. coli Aminopeptidase A	
1KPB	GroEL	
1LB3	Mouse L chain ferritin	
1QO1	Rotary Motor in ATP Synthase	
1VPX	Transaldolase	
1VRG	Propionyl-CoA carboxylase	++-
1W6T	Octameric Enolase	
1YG6	ClpP	
2AWB	Bacterial ribosome	
2BO9	Human carboxypeptidase A4	
2BYU	M.tuberculosis Acr1(Hsp 16.3)	
2GHO	Thermus aquaticus RNA polymerase	
2GLS	Glutamine Synthetase	
2H12	Acetobacter aceti citrate synthase	
2IDB	3-octaprenyl-4-hydroxybenzoate decarboxylase	
2REC	RECA hexamer	
3DY4	Yeast 20S proteasome	

Table S1. The experimental macromolecular complexes used for tomogram simulation and semantic segmentation.

S2 Implementation Details

All models were trained and tested on Keras[5] with Tensorflow[6] as the backend. The EMAN2 library [7] is used for simulating tomograms. The experiments were performed on a computer with three Nvidia GTX 1080 GPUs, one Intel Core i7-6800K CPU and 128GB memory.

References

[1] Pei, L., Xu, M., Frazier, Z., Alber, F.: Simulating cryo electron tomograms of crowded cell cytoplasm for assessment of automated particle picking. BMC bioinformatics **17**(1) (2016) 405

[2] Berman, H.M., Bhat, T.N., Bourne, P.E., Feng, Z., Gilliland, G., Weissig, H., Westbrook, J.: The protein data bank and the challenge of structural genomics. Nature Structural & Molecular Biology **7** (2000) 957–959
[3] Guo, Q., Lehmer, C., Martínez-Sánchez, A., Rudack, T., Beck, F., Hartmann, H., Pérez-Berlanga, M., Frottin, F., Hipp, M., Hartl, U., Edbauer, D., Baumeister, W., Fernández-Busnadiego, R.: In Situ Structure of Neuronal C9ORF72 Poly-GA Aggregates Reveals Proteasome Recruitment. Cell doi:10.1016/j.cell.2017.12.030 (2018)
[4] Zeng, X., Leung, M.R., Zeev-Ben-Mordehai, T., Xu, M.: A convolutional autoencoder approach for mining features in cellular electron cryo-tomograms and weakly supervised coarse segmentation. Journal of structural biology (2017)
[5] Chollet, F., et al.: Keras (2015)
[6] Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al.: Tensorflow: A system for large-scale machine learning. In: OSDI. Volume 16. (2016) 265–283
[7] Galaz-Montoya, J.G., Flanagan, J., Schmid, M.F., Ludtke, S.J.: Single particle tomography in eman2. Journal of structural biology **190**(3) (2015) 279–290